IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2018, accepted September 12, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2870528

Statistical Leakage Analysis Using Gaussian Mixture Model

HYUNJEONG KWON[1], (Student Member, IEEE), MINGYU WOO[2], (Student Member, IEEE), YOUNG HWAN KIM[1], (Senior Member, IEEE), AND SEOKHYEONG KANG[1], (Member, IEEE)

[1]Department of Electrical Engineering, Pohang University of Science and Technology, Pohang 37673, South Korea
[2]Department of Computer Engineering, University of California at San Diego, La Jolla, CA 92093, USA

Corresponding author: Seokhyeong Kang (shkang@postech.ac.kr)

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the "ICT Consilience Creative program" (IITP-2018-2011-1-00783) supervised by the IITP (Institute for Information & communications Technology Promotion).

ABSTRACT In the design process of advanced semiconductor devices, statistical leakage analysis has emerged as a major step due to uncertainties in the leakage current caused by the process variations. In this paper, a novel statistical leakage analysis which uses Gaussian mixture model (GMM) as the density function of leakage current is proposed. To estimate the probability density function, our proposed method clusters the rapidly converged leakage data using the GMM. The GMM can represent any distributions, so it is suitable to estimate the leakage distribution, which varies as the technology node or operating condition changes. In addition, our proposed method (SLA-GMM) defines a terminating condition that guarantees the convergence of the leakage data and prevents the underfitting or overfitting in the GMM modeling process. With sequential addition, SLA-GMM significantly reduced the error that can occur during the addition process. In studies with a goodness-of-fit test, SLA-GMM achieved up to 98% and 94% improvements in the Chi-square static and the K-S static compared with the previous method based on an analytic model.

INDEX TERMS Expectation-maximization algorithm, Gaussian mixture model, machine learning, statistical leakage analysis.

I. INTRODUCTION

Low power usage is mobile computing devices. To realize various strategies to reduce power use, the power must be accurately estimated during the design stage. Especially, leakage power is increasingly dominant in many applications [1]–[4]. Meanwhile, the process variations increase as device size is scaled down [5]–[7]. Process variations lead to variations in leakage current I_L because it is directly dependent on process parameters. For this reason, variation-aware I_L analysis, called statistical leakage analysis (SLA), has become an important step in the process of designing low-power devices. The previous SLA researches can be classified into two main approaches; sampling-based methods and analytic model-based methods.

Sampling-based methods use extracted parameter samples to simulate I_L. To reduce the complexity of traditional Monte-Carlo (MC) simulation, efficient leakage models have been developed [8], [9] and the convergence time of sampling has been reduced [10]–[12]. Sampling-based methods store all leakage data or histograms to estimate the leakage distribution.

Quasi-MC (QMC) simulation is the most effective sampling-based method in circuit simulation [11], [12]. QMC empirically reduced the number of leakage simulations required to achieve convergence, compared to traditional MC simulation. However, the number of simulations is difficult to predict, because calculation of the error bound of the QMC samples is more complicated than in traditional MC simulation [13]. Hence, leakage analysis using the QMC requires additional experiments to determine the terminating conditions. In addition, the histogram from the QMC simulation cannot generate a precise leakage distribution. The bin size of a histogram is generally a function of the number of data and the degree of spread of the data [14]–[16], so reduction in the number of leakage data of the QMC can result in wide bins and reduce the simulation accuracy.

Analytic model-based methods use an analytic model to represent I_L. This method does not require numerous simulations, so it can generate feasible models quickly. The first-order model [17] is the most representative of these methods. The first-order model is based on the assumptions that the $\log(I_L)$ has linear relationships with process parameters, and

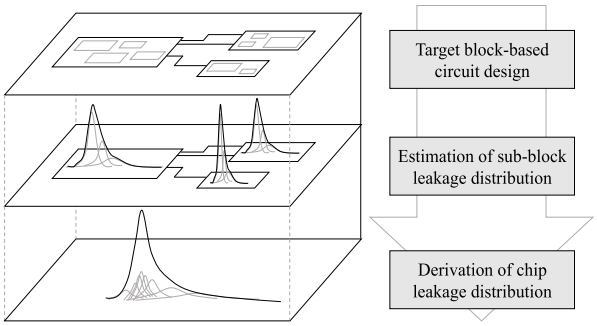

FIGURE 1. Illustration of our proposed method which obtains leakage distribution of a chip from sub-block leakage distributions.

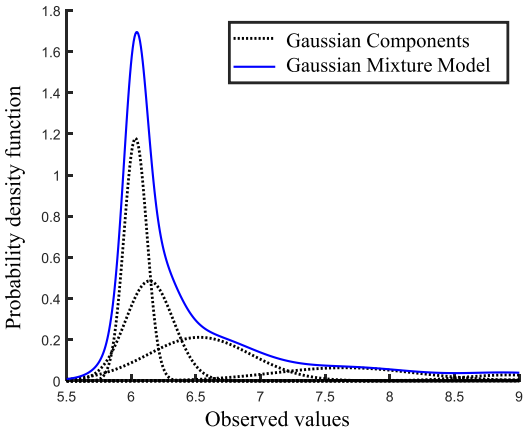

FIGURE 2. An example of simple Gaussian components and a Gaussian mixture model (GMM).

that the leakage follows a lognormal distribution. In advanced technology nodes, the generalized extreme value (GEV) distribution is used to improve the accuracy [18]. A proposed fourth-order model [19] is based on the assumption that the leakage distribution follows the GEV distribution.

The shape of leakage distribution varies as the technology node advances, and as operating conditions such as supply voltage change. Therefore, whenever the technology or operating condition changes, a new model must be developed to fit the leakage distribution to a well-known continuous function. For example, the accuracy of the first-order model [17] is significantly reduced in technologies that are more recent than BSIM4 models [20] due to their non-lognormality characteristics [21]. Furthermore, the GEV also cannot adapt to changes in the shape of the leakage distribution.

SLA cannot readily sum the leakage distributions of small circuits. The summation allows the leakage analysis at a smaller level than the full-chip level. Basically, the distribution of the sum of two arbitrary distributions can be obtained using the convolution of the two distributions. However, it is inefficient to perform a convolution every time the summation is performed and the convolution data points are changed. To avoid this, if the approximation to represent that the summation result using the same density function [18], [22], [23] is applied, the errors occurred in the summation process seriously degrade the accuracy.

Here, we propose a novel SLA method that accurately estimates the chip leakage distribution by using a Gaussian Mixture Model (GMM) (Fig. 1). To estimate the leakage distribution of a circuit, our proposed method (SLA-GMM) can represent any arbitrary function without fixing the shape of the leakage distribution to a particular continuous function. The terminating condition of SLA-GMM is defined using the error bound of Gaussian function. Finally, the summation in SLA-GMM does not require a convolution process, because the summation result can be represented as another GMM by using only the parameters of GMMs without any approximation.

The contributions of SLA-GMM are:

- It can represent any leakage distribution to maintain the accuracy even when the technology node or the operating point changes.
- It uses a terminating condition of leakage simulation by calculating the required number of leakage data for the convergence. The terminating condition can prevent the underfitting or overfitting during GMM clustering.
- It uses a new summation process, which is a necessary operation in leakage analysis. The summation operation in our method represents results in the same form without any approximation, and therefore greatly reduces the error.

Section II explains the background related to GMM. Section III presents SLA-GMM in detail. Section IV verifies the existence and uniqueness of SLA-GMM. Section V presents experimental results to validate SLA-GMM. Section VI concludes.

II. PRELIMINARY

A. GAUSSIAN MIXTURE MODEL

A GMM is the weighted sum of N Gaussian components (Fig. 2):

$$p(x) = \sum_{i}^{N_{GMM}} \omega_i \cdot N(x|\mu_i, \sigma_i) = \sum_{i}^{N_{GMM}} \omega_i \cdot \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \tag{1}$$

where μ_i is the mean, σ_i is the standard deviation, and ω_i is the weight of the i^{th} Gaussian component. Because the GMM consists of several Gaussian functions, it can represent various classes of continuous functions with reasonable accuracy. Hence the GMM can be used to estimate the probability density function (PDF) of the observed data. To utilize the GMM to estimate the distribution, the data must be clustered with several Gaussian components. The parameters of GMM such as N_{GMM}, μ_i, σ_i, and ω_i are determined during the clustering process. The most frequently-used method for GMM clustering is the expectation-maximization (EM) algorithm.

B. EXPECTATION-MAXIMIZATION ALGORITHM

EM algorithm iteratively finds the maximum likelihood estimates of the parameters of statistical models. It performs an E-step and an M-step iteratively. In the E-step, a function for likelihood is created using current parameter estimates. In the M-step, the parameters are updated by maximizing the likelihood function generated in the E-step. If the EM algorithm is used for clustering in GMM modeling, the following E- and M-steps are performed iteratively.

1) **E-step**: For each data point and Gaussian component, a membership weight $\alpha_{i,k}$ is calculated. $\alpha_{i,k}$ is the probability that the i^{th} data point belongs to the k^{th} cluster C_k. When N is the number of data and K is the number of clusters, the result of this step is an $N \times K$ matrix of membership weights where the elements sum to one in each row.
2) **M-step**: Using the data and the matrix of membership weights, a new parameter set is determined as (2).

$$\begin{aligned} \omega_k^{new} &= \frac{N_k}{N} \\ \mu_k^{new} &= \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} \alpha_{ik} \cdot x_i \\ \sigma^{2}{}_k^{new} &= \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} \alpha_{ik} \cdot \left(x_i - \mu_k^{new}\right)^2 \end{aligned} \tag{2}$$

where $N_k = \sum_{i=1}^{N} \alpha_{ik}$ is the effective number of data that are included in the k^{th} Gaussian component.

III. STATISTICAL LEAKAGE ESTIMATION USING GAUSSIAN MIXTURE MODEL

SLA-GMM estimates the leakage distribution of sub-block circuits and adds more than two leakage distributions sequentially. SLA-GMM consists of two main parts (Fig. 3). The first part is GMM modeling to estimate the leakage distribution. This step generates the leakage data for the GMM modeling, then after clustering, calculates the required number of leakage data. If the current number of data is less than the required number to ensure convergence, then additional leakage data are generated. The second part is sequential addition, which is essential in the SLA and helps to improve the scalability of SLA-GMM. SLA-GMM adds multiple leakage distributions sequentially, using the GMM parameters. This step can be useful to obtain the leakage distribution of a huge system that consists of many sub-blocks.

A. ESTIMATION OF THE LEAKAGE DISTRIBUTION USING GMM

In the SLA-GMM, we use three steps to estimate the leakage distribution. (1) We utilize the rapidly-converged samples to reduce the required number of simulation data. To achieve this goal, we use a Sobol sequence [24] for parameter sampling instead of pseudorandom numbers. (2) We use the GMM as the leakage distribution model to compensate for the loss of accuracy caused by use of the reduced number of simulation data. (6) We use a new method that calculates the number of simulation data required for convergence, and dynamically supplement the additional data if necessary. The calculated number of simulations can be used as a terminating condition and is useful to prevent overfitting or underfitting when estimating the parameters of the GMM.

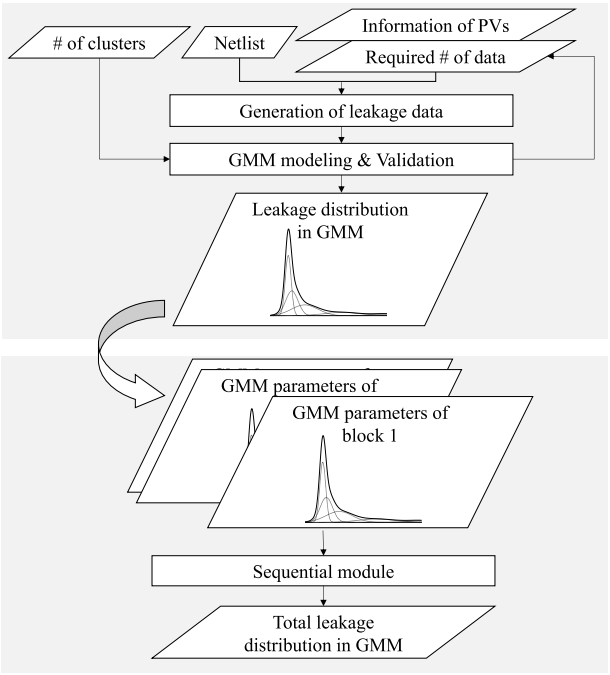

FIGURE 3. Overall flow of our statistical leakage analysis.

The estimation steps of the leakage distribution consist of three processes (Fig. 4). (1) Leakage data are generated for the GMM modeling, and clustering is performed. (2) The clustering is validated using a certain index. (6) After optimal clustering is performed, the required number of data is calculated and compared with the current number of data. If the current number of data is insufficient, leakage data for modeling are added. Details follow.

1) GENERATION OF LEAKAGE DATA FOR GMM MODELING

To generate leakage data to be used for the GMM modeling, parameter samples for the leakage simulation are extracted from the parameter spaces. The process parameters are assumed to follow Gaussian distributions. Therefore, after uniformly-distributed samples are extracted, we use the Box-Muller method to transform them to normally-distributed samples.

The purpose of SLA-GMM is to reduce the number of MC runs for efficiency. Therefore, the parameter samples should be extracted with a high convergence rate. SLA-GMM uses the Sobol quasi random sequence [24] rather than pseudorandom numbers as the uniform samples. The Sobol sequence converges faster than the pseudo random numbers because

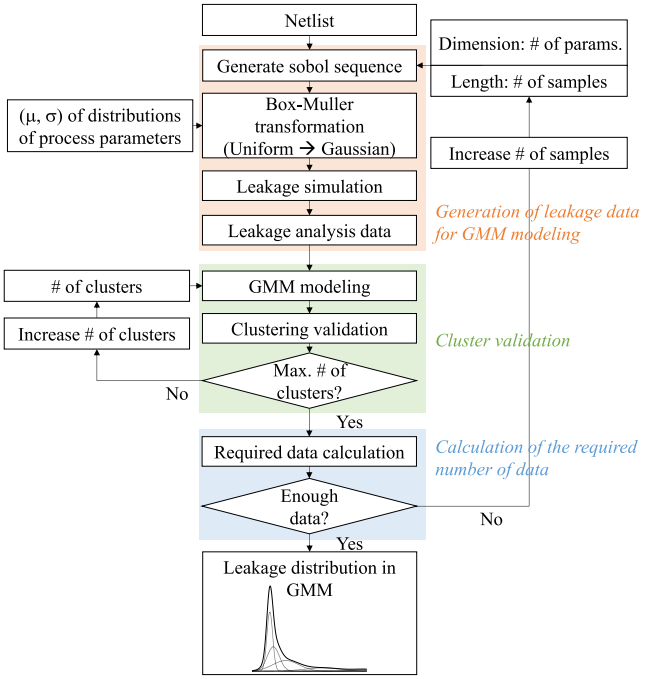

FIGURE 4. Algorithm flow of our proposed method which estimates the leakage distributions.

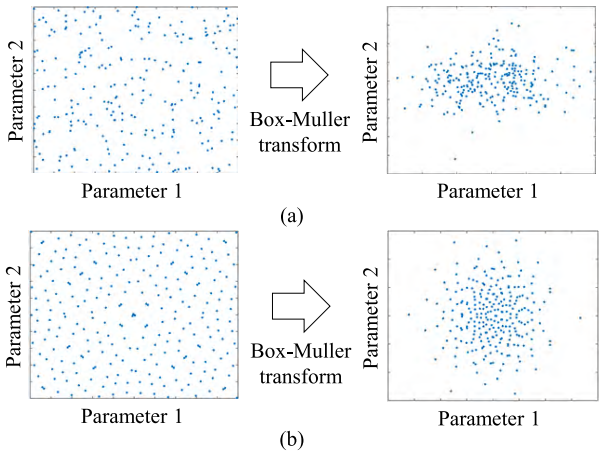

FIGURE 5. 256 samples transformed from (a) the pseudo random numbers and (b) the Sobol sequence.

it is determined by considering the discrepancy, which is a mathematical quantity that represents the non-uniformity of the points. Transformation using the Sobol sequence is closer to Gaussian than is transformation using pseudorandom numbers (Fig. 5).

Clustering is performed on leakage data obtained using any leakage model and the parameter samples. Then the quality of clustering is evaluated in two ways. (1) Cluster validation is performed to evaluate the degree of density and separation of data. (2) The required number of data is calculated. If the required number of data is larger than the current number of data, additional data are collected. Details follow.

2) CLSUTER VALIDATION

The clustering result is first validated using cluster cohesion, which evaluates the tightness of the cluster of leakage data; the measure is the sum of squares within cluster (*SSW*)

$$SSW = \sum_i \sum_{x \in C_i} (x - m_i)^2 \tag{3}$$

where x is a datum in the i^{th} cluster C_i and m_i is the mean of the data in C_i.

Then the separation of clusters is calculated as the sum of squares between clusters (*SSB*)

$$SSB = \sum_i |C_i|(m - m_i)^2 \tag{4}$$

where $|C_i|$ is the number of data in C_i and m is the mean of all data.

To optimize the number κ of clusters, we used the WB index [25].

$$WB\_index = \kappa \cdot \frac{SSW}{SSB} \tag{5}$$

The index is calculated sequentially. Usually, κ is initialized to two [26], then increased and WB-index is calculated each time. After the maximum κ ($\sqrt{L/2}$ as a rule of thumb [27]) is reached, κ that fields the smallest *WB_index* is selected as the optimal κ.

3) CALCULATION OF THE REQUIRED NUMBER OF DATA

To define the terminating condition and to prevent underfitting or overfitting, the number of required data should be calculated. The calculation for the convergence is simplified when the data are extracted from Gaussian distributions. SLA-GMM defines the required number of leakage data by applying the error for data obtained from Gaussian distribution as follows.

The required number of data from a Gaussian distribution with mean $\bar{x}$ and standard deviation S_x can be determined as [28]

$$N = \left[\frac{100 z_C S_x}{E\bar{x}}\right]^2 \tag{6}$$

where E is the allowed percentage error, z_C is the confidence level coefficient. If $z_C = 1.96$ and $E = 1$, the meaning of (6) is that using N data, we are 95% confident that the Gaussian distribution fitted from the data does not differ by more than 1% from the real Gaussian distribution.

Each component of the GMM is a Gaussian distribution and its clustered data, so each cluster must have more data than N in (6). Therefore, after leakage data are obtained and GMM modeling is completed with optimal κ, we test whether the numbers of data in all clusters satisfy (6). If this condition is satisfied, the data of each cluster can be regarded with z_C confidence level to come from a Gaussian distribution that is within E error. However, if one of the clusters does not satisfy the condition, additional data for GMM modeling are added. This step exploits the increase in accuracy of a

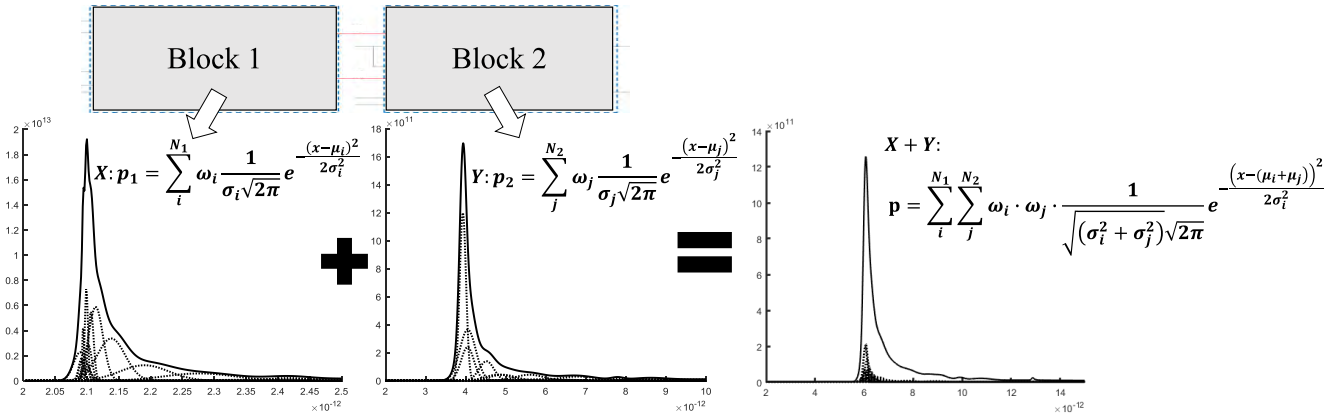

FIGURE 6. Total leakage distribution by adding two leakage distributions represented in GMM.

Sobol sequence as samples are added while fully reusing the existing sequences.

When cluster validation is finished and the required number of data is satisfied, the GMM represents the distribution of the leakage data. The next step is the sequential addition of two or more leakage distributions represented in the GMMs.

B. SEQUENTIAL ADDITION OF THE LEAKAGE DISTRIBUTION

The separately designed and analyzed data of sub-systems can be summed to estimate the full-chip leakage distribution (Fig. 6). Block-based design is common in modern designs, so the SLA should be applied to the sub-blocks in parallel, and the analysis results for the SLA should be applied to the full chip.

For this reason, we propose to use summation of the leakage distributions of the sub-systems to estimate the full-chip leakage distribution. Generally, the modeling errors before the summation operation are accumulated as summation errors. Therefore, the accuracy of estimating the leakage distribution degrades with each summation operation. This error is exacerbated when approximations must be used to allow sequential addition. The sum of two distributions represented in the GMMs can be also expressed in the GMM without any approximation, so SLA-GMM extremely reduces the rate of error accumulation during the summation step.

Let X and Y be random variables of I_{L} in two subsystems, and $Z = X + Y$ be a random variable of I_{L} of the total system. The PDF $f_{\mathrm{z}}(\mathrm{z})$ of Z can be generally obtained by convoluting the PDFs $f_{\mathrm{x}}(\mathrm{z})$ of X and $f_{\mathrm{y}}(\mathrm{z})$ of Y as

$$\begin{aligned} f_Z(z) &= f_X(z) * f_Y(z) \\ &= \int f_X(z-y)\cdot f_Y(y)\,dy \end{aligned} \tag{7}$$

For a Gaussian distribution, $f_{\mathrm{z}}(\mathrm{z})$ can be simply summarized after (7) is solved. The complexity of this calculation is greatly reduced using (8).

$$\begin{aligned} X: N \sim (\mu_1, \sigma_1^2), \quad Y: N \sim (\mu_2, \sigma_2^2) \\ \rightarrow Z: N \sim (\mu_1+\mu_2, \sigma_1^2+\sigma_2^2) \end{aligned} \tag{8}$$

SLA-GMM uses GMM as the leakage distribution, so

$$\begin{aligned} X: p_1 &= \sum_j^{N_1} \omega_j \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} \\ Y: p_2 &= \sum_k^{N_2} \omega_k \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \end{aligned} \tag{9}$$

We will use the characteristic function

$$\varphi_{\mathrm{X}} = \mathrm{E}\left[e^{itx}\right] \tag{10}$$

and the fact that the characteristic function of the sum of two random variables is the product of their characteristic functions. Using (10), the characteristic function of X is

$$\begin{aligned} \varphi_{\mathrm{X}} &= \mathrm{E}\left[e^{itx}\right] = \int_{-\infty}^{\infty} p_1 e^{itx}\,dx \\ &= \int_{-\infty}^{\infty} \sum_j^{N_1} \omega_j \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} e^{itx}\,dx \\ &= \sum_j^{N_1} \omega_j \int_{-\infty}^{\infty} \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} e^{itx}\,dx \end{aligned} \tag{11}$$

The characteristic function of a Gaussian component with the mean μ and variance σ^2 is $\exp(it\mu + \sigma^2 i^2 t^2/2)$, so (11) can be simplified to

$$\varphi_{\mathrm{X}} = \sum_j^{N_1} \omega_j e^{it\mu_j - \frac{t^2\sigma_j^2}{2}}, \tag{12}$$

and the characteristic function of Y can be represented as

$$\varphi_{\mathrm{Y}} = \sum_k^{N_2} \omega_k e^{it\mu_k - \frac{t^2\sigma_k^2}{2}}. \tag{13}$$

The characteristic function of $X+Y$ is equal to the product of the characteristic functions of X and Y:

$$\varphi_{\mathrm{X+Y}} = \varphi_{\mathrm{X}}\varphi_{\mathrm{Y}} = \sum_j^{N_1}\sum_k^{N_2} \omega_j \cdot \omega_k e^{it(\mu_j+\mu_k) - \frac{t^2(\sigma_j^2+\sigma_k^2)}{2}}, \tag{14}$$

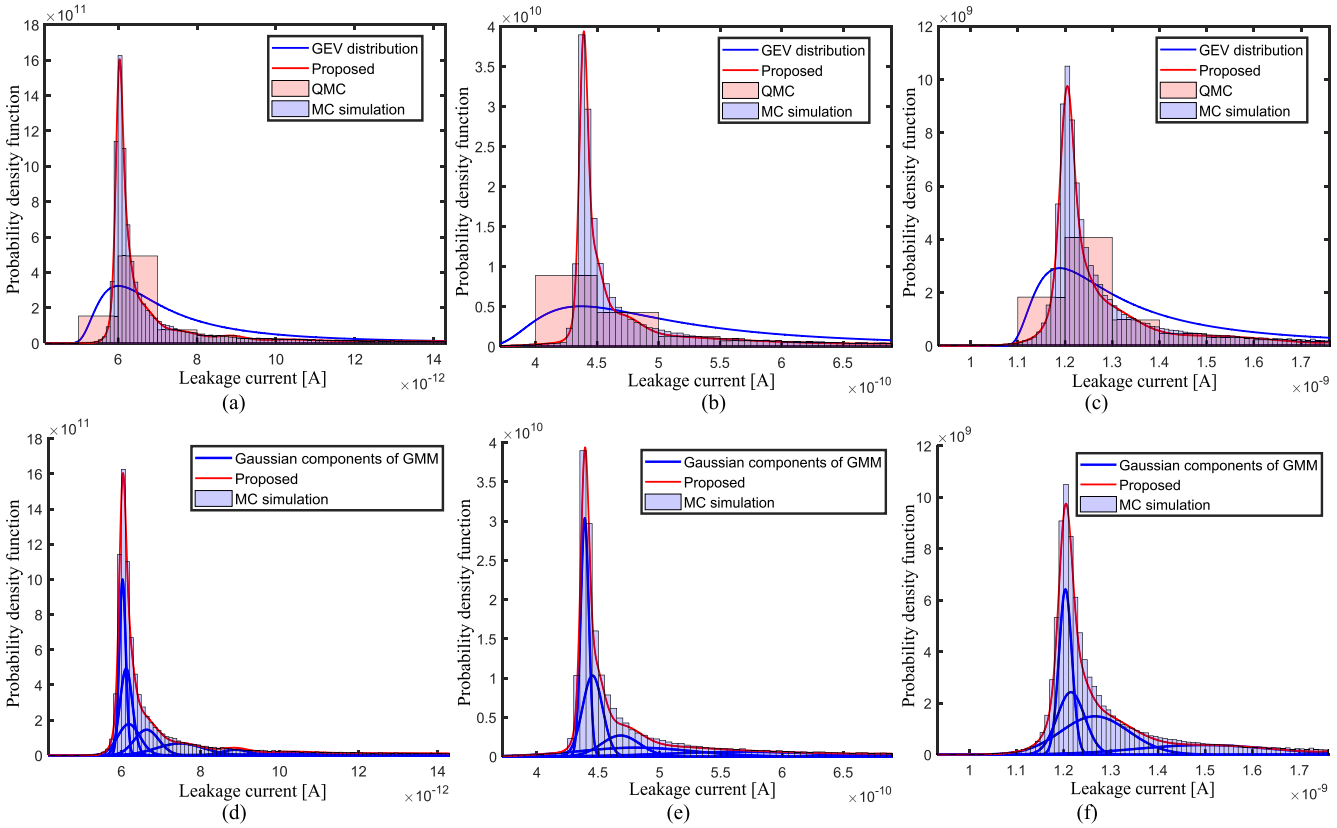

FIGURE 7. Accuracy comparison of the proposed method with other benchmark methods when estimating the leakage distribution of (a) c17, (b) c1355, and (c) c3540. Gaussian components and Gaussian mixture model which represents the leakage distribution of (d) c17, (e) c1355, and (f) c3540.

which implies that the PDF of Z is

$$Z : p_3 = \sum_{j}^{N_1} \sum_{k}^{N_2} \omega_j \cdot \omega_k \frac{1}{\sqrt{\left(\sigma_j^2 + \sigma_k^2\right)}\sqrt{2\pi}} e^{-\frac{\left(x-\left(\mu_j+\mu_k\right)\right)^2}{2\left(\sigma_j^2+\sigma_k^2\right)}} \tag{15}$$

Equation (15) means that the density function of Z is represented using Gaussian pairs from the GMMs of X and Y. Each Gaussian component of X forms a pair with a Gaussian component of Y. Each pair of Gaussian components generates a new Gaussian component of Z. The mean and variance of the newly- generated Gaussian component are respectively the sum of mean and variance of the Gaussians the pair. The weight of the new Gaussian component is the product of the weights of the Gaussians of the pairs. The new Gaussians of all pairs from X and Y form the GMM for Z. By exploiting these characteristics, the addition can be simply implemented in SLA-GMM without any approximation. This advantage slows the accumulation of error during the summation step.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL ENVIRONMENT

The proposed leakage analysis method was compared with existing methods. As a benchmark method, we used the most recent sampling-based method [12] which is the most effective smart sampling [11]. We also compared the state-of-the-art analytic model-based method, which uses the generalized extreme value distribution [18]. One hundred thousand MC data were used as the basis for evaluation of the accuracy of SLA-GMM and the benchmark methods.

In the following experiments, ''16-nm predictive technology model high-performance'' was used as the transistor model [29]. Variations in gate length, width, oxide thickness, and threshold voltage were considered. The 3σ values of the process parameter distributions were set to 18% of their mean values. For all benchmark methods, the BSIM4 model [20] was used as the leakage model, and HSPICE [30] was used as the leakage simulator. This is to exclude the impact on the accuracy of the leakage model when comparing the accuracy of each method. Ten ISCAS 85 circuits were used as the benchmark circuits.

Gaussian mixture modeling using the EM algorithm was implemented in Python language, and was based on open sources related to the machine learning algorithms provided by Tensorflow [31]. The other processes of SLA-GMM and the benchmark methods were implemented in C++. All implementations were done on Intel(R) Xeon(R) CPU E5-2690 @ 2.90 GHz.

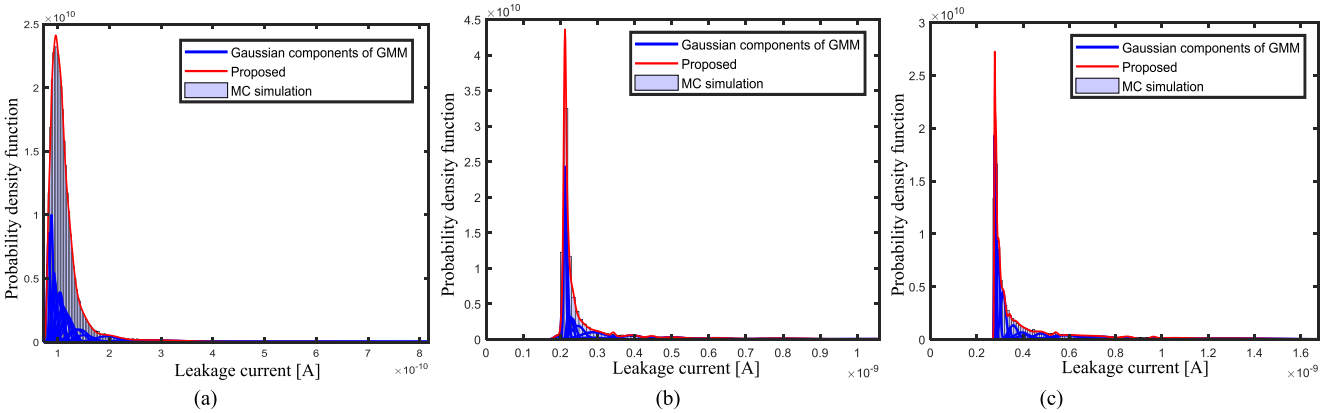

FIGURE 8. The changes of leakage distribution according to the supply voltage. The proposed method can cover the change of the shape of the leakage distribution at (a) low, (b) nominal, and (c) high supply voltages.

TABLE 1. Goodness-of-fit test of the proposed method and the benchmark methods with MC simulation.

Circuit	Chi-square statistic χ^2			K-S statistic D		
	GEV	QMC	Proposed	GEV	QMC	Proposed
c17	7.69 x 10^{12}	9.53 x 10^{12}	**1.23 x 10^{11}**	0.198	0.00596	**0.00464**
c432	1.54 x 10^{11}	1.68 x 10^{10}	**2.52 x 10^{9}**	0.188	0.0063	**0.00434**
c499	1.05 x 10^{11}	1.68 x 10^{10}	**1.42 x 10^{9}**	0.161	0.00995	**0.00671**
c880	8.16 x 10^{10}	3.05 x 10^{10}	**1.18 x 10^{9}**	0.173	0.00689	**0.00561**
c1355	9.49 x 10^{10}	2.40 x 10^{11}	**7.10 x 10^{9}**	0.204	0.0092	**0.0089**
c1908	3.70 x 10^{10}	1.20 x 10^{11}	**1.95 x 10^{9}**	0.143	**0.00679**	0.0296
c2670	3.21 x 10^{10}	6.54 x 10^{10}	**1.02 x 10^{9}**	0.146	**0.00813**	0.0136
c3540	2.20 x 10^{10}	1.98 x 10^{10}	**8.94 x 10^{8}**	0.145	0.01	**0.00622**
c5315	1.30 x 10^{10}	6.93 x 10^{9}	**8.15 x 10^{8}**	0.147	0.012	**0.0102**
c6288	3.60 x 10^{10}	1.25 x 10^{9}	**2.51 x 10^{8}**	0.189	0.0144	**0.0106**
c7552	4.68 x 10^{9}	2.01 x 10^{10}	**6.27 x 10^{8}**	0.151	0.013	**0.00991**
Avg. normalized error	1.00	1.22	**0.0170**	1.00	**0.0556**	0.0598

B. VALIDATION OF THE ESTIMATION OF THE LEAKAGE DISTRIBUTION USING GMM

1) ACCURACY COMPARISON

We used goodness-of-fit statistics to quantify the accuracy of SLA-GMM, and verified the WB index by comparing the estimated κ with the actual optimal number of clusters.

Firstly, the PDFs of the golden MC simulation, Quasi-MC (QMC), the GEV distribution, and SLA-GMM were graphically compared (Fig. 7a-c). The required number of QMC data is not defined, so the same number of data used for SLA-GMM was used for the QMC simulation and for fitting to the GEV distribution. The bin size of the QMC histogram was very large as expected because the bin size is proportional to the number of data. Therefore, even though the QMC helps improve the convergence rate of the leakage data, the QMC did not accurately estimate the PDF. Also, the GEV distribution was far from the actual leakage distribution, even though the GEV distribution has been known as the most accurate continuous function to approximate the leakage distribution. This failure occurred because the actual leakage distribution changes as the technology or supply voltage varies. Therefore, a fixed continuous function is not appropriate to estimate the leakage distribution.

The PDF of SLA-GMM consists of several Gaussian components (Fig. 7d-f). Thus, SLA-GMM can represent any functions, unlike a method that uses on class of continuous function. Therefore, even if the technology or the operating conditions such as supply voltage change, SLA-GMM can estimate the leakage distribution adaptively.

The shape of leakage distribution changes as the supply voltage changes (Fig. 8). However, the existing methods based on analytic models assume that the leakage distribution follows a certain fixed distribution function, and therefore they cannot estimate leakage distribution whose shape changes; in contrast SLA-GMM is resilient to changes of the shape of leakage distribution because the GMM can represent any continuous function. Under high or low supply voltage, SLA-GMM estimated the leakage distribution accurately (Fig. 8).

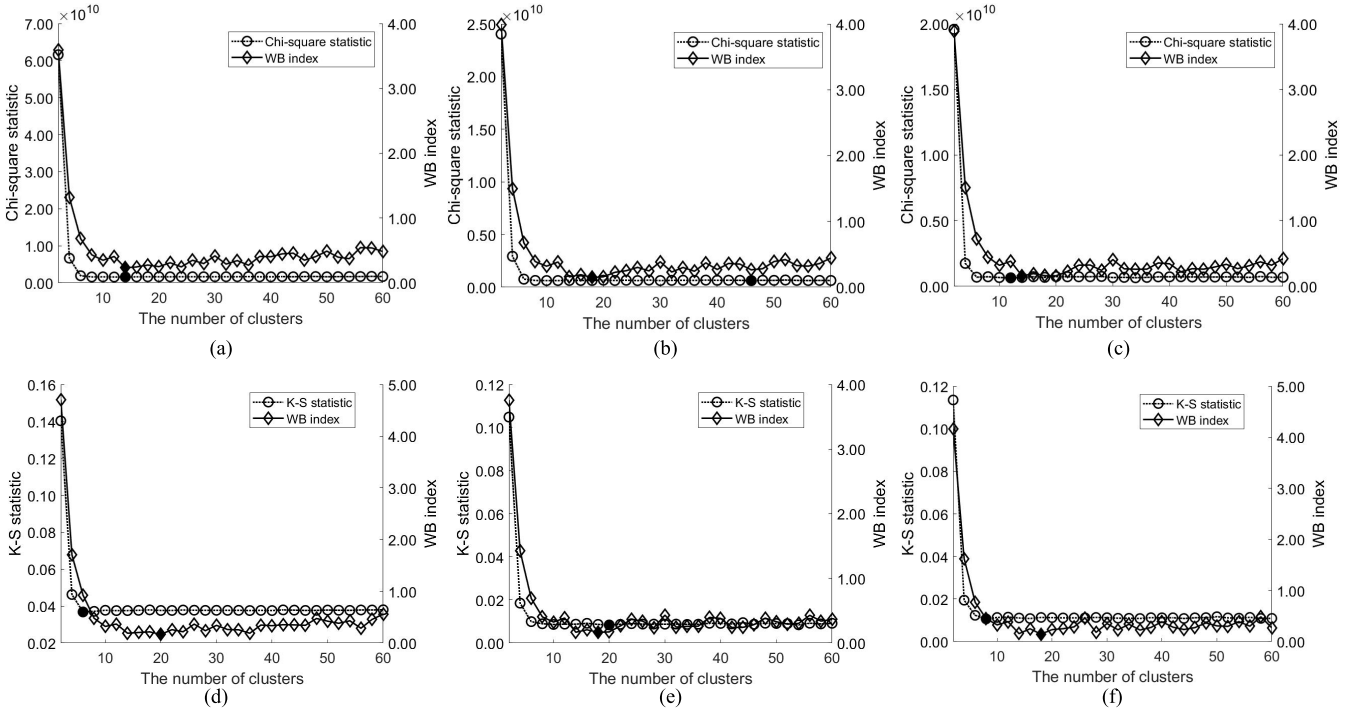

FIGURE 9. Chi-square statistic ((a)-(c)) and K-S statistic ((d)-(f)) with WB index in terms of the number of clusters in order to validate WB index for determining the optimal number of clusters using (a) c499, (b) c2670, (c) c5315, (d) c1355, (e) c3540, and (f) c7552 circuits.

TABLE 2. Chi-square statistic and K-S statistic when the optimal number of clusters determined by WB index is used are compared with those when the real optimal number of clusters is used.

Circuit	Chi-square statistic					K-S statistic				
	# of clusters (A)	Relative χ^2 at A	# of clusters (B)	Relative χ^2 at B	Diff. [%]	# of clusters (A)	Relative D at A	# of clusters (C)	Relative D at C	Diff. [%]
c17	9	0.0456	5	0.0412	0.441	9	0.0312	32	0.0295	0.164
c432	9	0.0352	6	0.0329	0.235	9	0.0826	23	0.079	0.356
c499	7	0.0257	7	0.0257	0.000	7	0.0667	18	0.0637	0.301
c880	12	0.0317	16	0.0308	0.089	12	0.103	8	0.103	0.049
c1355	10	0.0498	5	0.0474	0.236	10	0.268	3	0.262	0.571
c1908	9	0.0344	5	0.033	0.136	9	0.251	18	0.248	0.223
c2670	9	0.0266	23	0.0259	0.072	9	0.141	5	0.138	0.299
c3540	9	0.0343	6	0.0321	0.219	9	0.0823	10	0.0799	0.246
c5315	7	0.0346	6	0.0329	0.163	7	0.0832	9	0.0811	0.207
c6288	8	0.0775	26	0.0605	1.70	8	0.357	4	0.282	7.45
c7552	9	0.0494	29	0.0449	0.448	9	0.0998	4	0.0938	0.596

The Chi-square statistic χ^2 and the Kolmogorov-Smirnov (K-S) statistic D were used to quantify the similarity between two distributions by comparing the PDF and cumulative density function (CDF), respectively [10] (Table I). Small values of χ^2 and K-S D indicate accurately-estimated distributions. Average χ^2 of SLA-GMM was 1.28×10^{10}, which is 98.3% and 98.6% smaller than those of GEV and QMC, respectively. Similarly, the average K-S D of SLA-GMM was 0.01, which is 94.0% smaller than that of GEV, and comparable to that of QMC. These results showed that SLA-GMM is the most accurate of these methods to estimate the PDF and the CDF. K-S D of our method is expected to be much smaller than that of the QMC after the summation operation, because χ^2 has a large influence on the accuracy of the summation step. This effect will be described in the next section.

We chose the WB index as the cluster validation index. In our experiments, the trend of the WB index followed

TABLE 3. Total runtime for benchmark methods and the proposed method.

Circuit	Collecting data (A) [s]	GMM modeling					MC [s]
		Clustering [s]	Cluster validation [s]	Calculation of required # of data [s]	Total (B) [s]	Overhead (B/A) [%]	
c17	698	454	1.44	0.00742	457	65.5	20,016
c432	3,888	2,866	17.2	0.0374	2,884	74.2	65,160
c499	3,816	490	1.67	0.0125	490	12.8	108,720
c880	5,436	1,130	4.64	0.0164	1,134	20.9	99,000
c1355	6,264	731	2.72	0.0156	734	11.7	138,960
c1908	5,364	446	1.47	0.00842	450	8.39	153,720
c2670	7,092	450	1.44	0.00760	450	6.35	202,320
c3540	7,056	243	0.623	0.00475	244	3.46	281,520
c5315	11,052	239	0.634	0.00619	240	2.17	442,800
c6288	35,388	1,696	8.35	0.0544	1,703	4.81	504,000
c7552	14,220	239	0.626	0.00742	240	1.69	568,800

TABLE 4. Results of K-S statistic and runtime profiling of the sequential addition.

Circuit	# of sub-blocks	QMC				Proposed				Overhead (B-A)/A [%]
		K-S statistic	Estimation [s]	Summation [s]	Total (A) [s]	K-S statistic	Estimation [s]	Summation [s]	Total (B) [s]	
C5315	0	0.012	11,052	-	11,052	0.010	11,292	-	11,292	2.17
	2	0.834	7,500	0.0458	7,500	0.179	7,740	0.0005	7,740	3.20
	4	0.876	5,040	0.148	5,040	0.346	5,346	1.08	5,347	6.09
	8	0.901	2,160	0.306	2,160	0.496	2,178	9.10	2,187	1.24
C6288	0	0.014	35,388	-	35,388	0.011	37,091	-	37,091	4.81
	2	0.688	8598	0.0603	8,598	0.129	8664	0.0004	8,664	0.77
	4	0.813	5523	0.171	5,523	0.300	5657	0.0003	5,657	2.42
	8	0.992	2758	1.069	2,759	0.457	2817	9.36	2,826	2.42
C7552	0	0.013	14,220	-	14,220	0.009	14,460	-	14,460	1.69
	2	0.839	11,820	0.0404	11,820	0.183	12,060	0.0002	12,060	2.03
	4	0.893	6,660	0.105	6,660	0.357	6,900	1.08	6,901	3.62
	8	0.899	2,640	0.196	2,640	0.500	2,664	9.03	2,673	1.24

the trend of goodness-of-fit test results (Fig. 9). Small WB index, χ^2, and K-S D mean good estimation, so we can determine the number of clusters that reduce the goodness-of-fit statistics by using the smallest WB index value; i.e., we first use the WB index to select the number of clusters (filled circle, Fig. 9). This process yields clustering results that have goodness-of-fit that are not far from the real solution (filled diamond, Fig. 9)

We tabulated (Table 2) the numerical evaluation results of Fig. 9 for all benchmark circuits The optimal number of clusters determined by the WB index increased the statistics by $< 8\%$. Therefore, the WB index is effective to determine the optimal number of clusters when goodness-of-fit results are unknown.

2) RUNTIME

Runtime of SLA-GMM consists of data-collection time and GMM-modeling time. The number of QMC data was set as the same as that determined by SLA-GMM, so data-collecting time required by the QMC was equal. The GMM modeling time was divided into clustering, cluster validation, and time required to calculate the required number of data. Data clustering was performed repeatedly until optimal numbers of clusters and data were determined. Generally, the one-time clustering time was several seconds, and increased as the number of clusters and data increased.

Data-collecting time was greater than the GMM-modeling time (Table 3). The accurate results of SLA-GMM were achieved at the expense of 19.27 % average runtime overhead compared to the data-collecting time. This overhead of the GMM modeling decreased as the circuit size was increased, because the data-collecting time increases linearly as the circuit size increases, whereas the modeling time does not increase. Considering the increased accuracy of SLA-GMM (Table I), this overhead is less than the cost of collecting additional data.

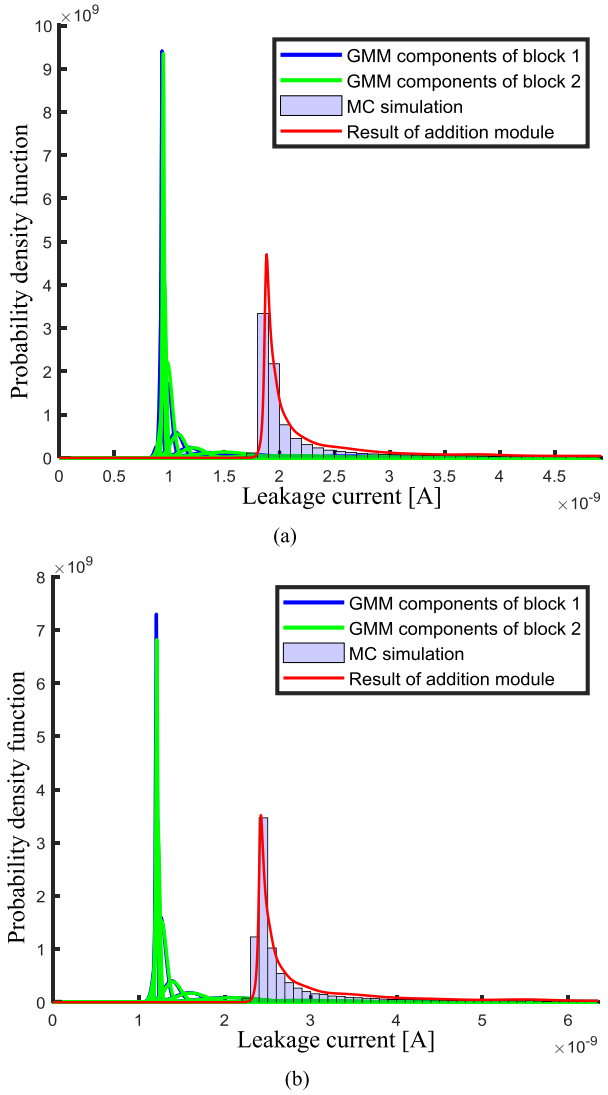

FIGURE 10. GMMs representing the leakage distribution of the sub-blocks and the result of the sequential addition module when using (a) c5315 and (b) c7552.

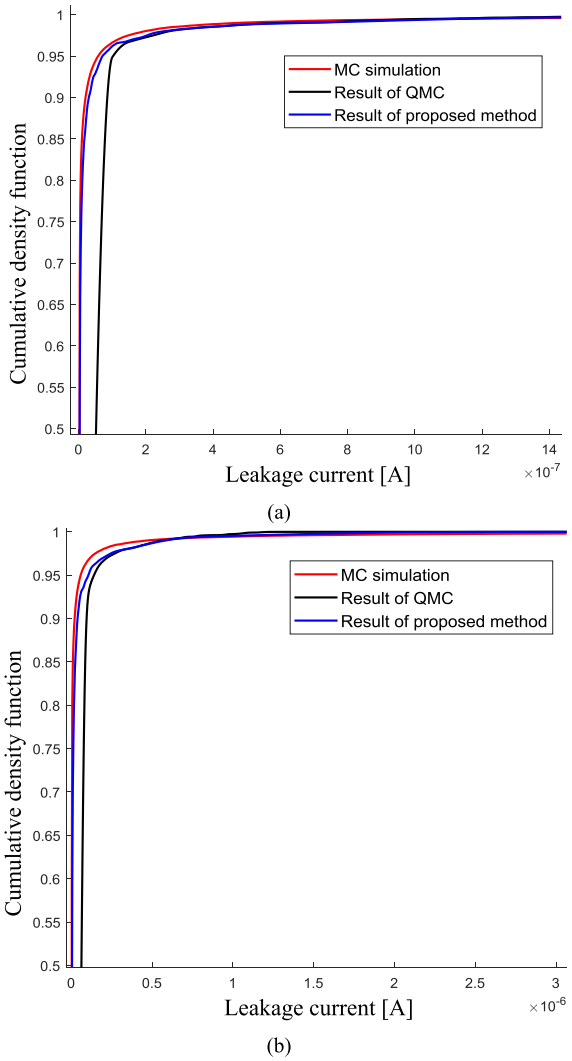

FIGURE 11. Cumulative density functions of the estimated leakage distributions of c7552 when the number of sub-blocks is (a) two and (b) four.

3) VALIDATION OF SEQUENTIAL ADDITION

SLA-GMM is effective to analyze the leakage distribution of a system that consists of several sub-blocks. To show the effectiveness of the sequential addition, we used a partitioning algorithm to break ISCAS 85 circuits into several virtual blocks.

We used the MLPart for the multi-level min-cut partitioning [32]. We used the Fiduccia-Mattheyses algorithm to perform the top-level partitioning, and the heavy-edge matching method for coarsening. We use V-cycling, which is composed of the repeated clustering-partitioning-refinement procedure, to use a solution that had been generated by a previous execution. We used a clustering ratio of 1.3, which has been shown empirically to be optimal [33].

We used the three largest ISCAS 85 circuits in this experiment. We first used the idea in Section III-A to estimate the leakage distributions of sub-blocks. Then we sequentially added the leakage distributions of the sub-blocks only using the parameters of the GMMs (Section III-B) to obtain Gaussian components of the GMMs of two sub-blocks and the resulting final leakage distribution of the total system (Fig. 10). The summation of the separately-modeled GMM distributions accurately estimated the full-chip leakage distribution. The CDF of SLA-GMM was much closer to the MC simulation result than was CDF of the QMC (Fig. 11).

The summation of the distributions is performed by the convolution of PDFs, so the error of estimating the PDF of a sub-block directly affects the summation of the PDF. For this reason, the QMC, which showed higher χ^2 than SLA-GMM (Table 1) showed abrupt increase in D (> 0.6) after just one summation, compared to D of SLA-DMM (Table 4). In addition, as the number of summation operations increases, the error of the estimate of the PDF increases so the accuracy

of the summation generally degrades. Due to the low error of SLA-GMM in each summation step, SLA-GMM slowed the rate of degradation of the accuracy compared to QMC, when leakage distributions of two or more sub-blocks were summed (Table 4).

We compared the runtime of the QMC and SLA-GMM including the sequential addition module (Table IV). We used the most common fast convolution algorithm which uses the fast Fourier transform. Also, we parallized the process to use the same number of cores as the number of sub-blocks. Although the runtime for the summation of the QMC was less than that of SLA-GMM, the summation of the QMC must be performed every time the data points of interest change. However, the summation results are represented in an analytic function of GMM in SLA-GMM, so it does not require additional operations even when the data points change. In addition, the results showed that SLA-GMM estimated the summation results with lower K-S D than QMC, which imposing only 2.64 % overhead on average, compared to QMC (Table IV). These results show that SLA-GMM has the high scalability and therefore can be used in SLA on a large circuit that consists of sub-blocks.

V. CONCLUSION

We proposed a method to use the Gaussian mixture model (GMM) in statistical leakage analysis (SLA). SLA-GMM consists of various shapes of leakage distribution modeling and sequential addition. The GMM can represent arbitrary functions and can be added sequentially with no assumptions, so it was considered to be a suitable model of leakage distribution. We also proposed a method based on the error bound of Gaussian distribution to calculate the required number of data. In experiments, SLA-GMM reduced χ^2 by 98% and K-S D by 94% compared to the GEV distribution. The advantage of accuracy was achieved with 19.27% runtime overhead on average, compared to generating leakage data. Experiments on the sequential addition module showed the scalability of SLA-GMM. SLA-GMM reduced the accuracy loss more than the QMC did. Using the incremental and parallel characteristics, SLA-GMM can for SLA analysis of large circuits that consist of sub-blocks.

REFERENCES

[1] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of CMOS," in *IEDM Tech. Dig.*, Dec. 2005, pp. 9–15.

[2] M. Seok, S. Hanson, D. Blaauw, and D. Sylvester, "Sleep mode analysis and optimization with minimal-sized power gating switch for ultralow V_{dd} operation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 4, pp. 605–615, Apr. 2012.

[3] C. Pratap and L. Kumre, "A new technique for leakage power reduction in CMOS circuit by using DSM," *Int. J. Comput. Appl.*, vol. 179, pp. 26–31, Sep. 2017.

[4] J. Zhan, J. Ouyang, F. Ge, J. Zhao, and Y. Xie, "DimNoC: A dim silicon approach towards power-efficient on-chip network," in *Proc. Design Autom. Conf. (DAC)*, Jun. 2015, pp. 1–6.

[5] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. Design Autom. Conf. (DAC)*, Jun. 2003, pp. 338–342.

[6] S. R. Nassif, "Design for variability in DSM technologies [deep submicron technologies]," in *Proc. Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2000, pp. 451–454.

[7] C. Visweswariah, "Death, taxes and failing chips," in *Proc. Design Autom. Conf. (DAC)*, Jun. 2003, pp. 343–347.

[8] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for VLSI: Timing and Power*. New York, NY, USA: Springer, 2005, pp. 42–45.

[9] J. Kim and Y. H. Kim, "Hybrid gate-level leakage model for Monte Carlo analysis on multiple GPUs," *IEEE Access*, vol. 2, pp. 183–194, Mar. 2014.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*. Cambridge, MA, USA: Cambridge Univ. Press, 2007, pp. 730–740.

[11] A. Singhee and R. A. Rutenbar, "Why quasi-Monte Carlo is better than Monte Carlo or Latin hypercube sampling for statistical circuit analysis," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 29, no. 11, pp. 1763–1776, Nov. 2010.

[12] V. Veetil, D. Sylvester, D. Blaauw, S. Shah, and S. Rochel, "Efficient smart sampling based full-chip leakage analysis for intra-die variation considering state dependence," in *Proc. Design Autom. Conf. (DAC)*, Jul. 2009, pp. 154–159.

[13] B. Tuffin, "Randomization of quasi-Monte Carlo methods for error estimation: Survey and normal approximation," *Monte Carlo Methods Appl.*, vol. 10, pp. 617–628, Dec. 2004.

[14] H. A. Sturges, "The choice of a class interval," *J. Amer. Stat. Assoc.*, vol. 21, pp. 65–66, Mar. 1926.

[15] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," *Probab. Theory Rel. Fields*, vol. 57, pp. 453–476, Dec. 1981.

[16] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[17] H. Chang and S. S. Sapatnekar, "Full-chip analysis of leakage power under process variations, including spatial correlations," in *Proc. Design Autom. Conf. (DAC)*, Jun. 2005, pp. 523–528.

[18] H. Aghababa, A. Khosropour, A. Afzali-Kusha, B. Forouzandeh, and M. Pedram, "Statistical estimation of leakage power dissipation in nanoscale complementary metal oxide semiconductor digital circuits using generalised extreme value distribution," *IET Circuits, Devices Syst.*, vol. 6, pp. 273–278, Sep. 2012.

[19] L. Cheng, P. Gupta, and L. He, "Efficient additive statistical leakage estimation," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 28, no. 11, pp. 1777–1781, Nov. 2009.

[20] T. H. Morshed *et al.* (Apr. 2009). *BSIM4.6.4 MOSFET Model—User's Manual*. [Online]. Available: http://bsim.berkeley.edu/models/bsim4/

[21] W. Kim, H. S. Park, D. J. Hyun, Y. H. Kim, and K. T. Do, "Investigation on the non-lognormal characteristic of the leakage current distribution under process variation," in *Proc. New Explor. Technol. (NEXT)*, Seoul, South Korea, 2007, pp. 126–132.

[22] N. C. Beaulieu, A. A. Abu-Dayya, and P. J. McLane, "Comparison of methods of computing lognormal sum distributions and outages for digital wireless applications," in *Proc. IEEE Int. Conf. Commun.*, May 1994, pp. 1270–1275.

[23] W. Kim, K. T. Do, and Y. H. Kim, "Statistical leakage estimation based on sequential addition of cell leakage currents," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 4, pp. 602–615, Apr. 2010.

[24] I. M. Sobol, "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Comput. Math. Math. Phys.*, vol. 7, pp. 86–112, May 1967.

[25] Q. Zhao and P. Fränti, "WB-index: A sum-of-squares based index for cluster validity," *Data Knowl. Eng.*, vol. 92, pp. 77–89, Jul. 2014.

[26] Q. Zhao, "Cluster validity in clustering methods," Ph.D. dissertation, Dept. Forestry Natural Sci., Univ. Eastern Finland, Kuopio, Finland, 2012.

[27] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. New York, NY, USA: Academic, 1979.

[28] M. Driels and Y. Shin, "Determining the number of iterations for Monte Carlo simulations of weapon effectiveness," Naval Postgraduate School, Monterey, CA, USA, Tech. Rep. NPS-MAE-04-005, 2004.

[29] (2012). *Predictive Technology Model*. [Online]. Available: http://ptm.asu.edu

[30] *HSPICE*, Synopsys, Mountain View, CA, USA, 2015.

[31] (2018). *Gaussian Mixture Clustering*. [Online]. Available: https://github.com/tensorflow/tensorflow/tree/master/tensorflow/contrib/factorization/python/ops
[32] D. A. Papa and I. L. Markov, *Handbook of Approximation Algorithms and Metaheuristics*. Boca Raton, FL, USA: CRC Press, 2006.
[33] A. E. Caldwell, A. B. Kahng, and I. L. Markov, "Improved algorithms for hypergraph bipartitioning," in *Proc. Asia South Pacific Design Automat. Conf. (ASPDAC)*, Jan. 2000, pp. 661–666.

HYUNJEONG KWON received the B.E. degree in electronic and electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electronic and electrical engineering. Her current research interests include variation-aware circuit analysis methodologies.

MINGYU WOO received the B.S. and M.S. degrees in electrical engineering from the Ulsan National Institute of Science and Technology, Ulsan, South Korea, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA, USA. His current research interests include design for manufacturing and physical design optimization.

YOUNG HWAN KIM (S'86–M'89–SM'14) received the B.E. degree in electronics from Kyungpook National University, South Korea, in 1977, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 1985 and 1988, respectively.

From 1977 to 1982, he was with the Agency for Defense Development, South Korea, where he was involved in various military research projects, including the development of auto-pilot guidance and control systems. From 1983 to 1988, he was a Post-Graduate Researcher with the Electronic Research Laboratory, University of California at Berkeley, where he was involved in developing VLSI CAD programs. He is currently a Professor with the Division of Electronic and Computer Engineering, Pohang University of Science and Technology, South Korea. His research interests include plasma and liquid crystal display systems, multimedia circuit design, MPSoC and GPGPU sys-tem design for display and computer vision applications, statistical analysis and design technology for deep-submicron semiconductor devices, and power noise analysis. He has served as an Editor for the *Journal of the Institute of Electronics Engineers of Korea*, and as the General Chair and a committee member of various Korean domestic and international technical conferences, including the International SoC Design Conference, the IEEE ISCAS 2012, and the IEEE APCCAS 2016.

SEOKHYEONG KANG (S'11–M'13) received the B.S. and M.S. degrees in electrical engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 1999 and 2001, respectively, and the Ph.D. degree from the VLSI CAD Laboratory, University of California at San Diego, La Jolla, USA, in 2013. He was with the System-on-Chip (SoC) Development Team, Samsung Electronics, Suwon, South Korea, from 2001 to 2008, where he was involved in development and commercialization of optical disk drive SoC.

He was a Professor with the Department of Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea, from 2014 to 2018. He has been a Professor with the Department of Electrical Engineering, Pohang University of Science and Technology, since 2018. His current research interests include low-power design optimization and cost-driven methodology for chip implementation.